גלעד עיני 034744920

מתן פינץ 300895315

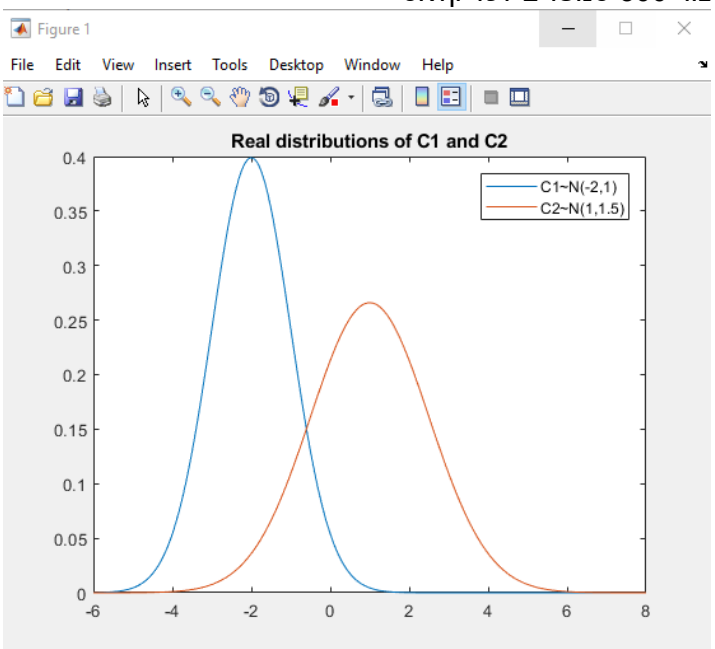
למידה ממוכנת

תרגיל בית מספר 2

1. יהי C1~N(-2,1) ו- C2~N(1,1.5) וגם P(C1)=P(C2)

- הערה: לא שמרנו את הדאטה. לכן בכל הרצה הנתונים משתנים קלות והערכים למטה נכונים רק להרצה הזאת ספציפית אך כן אמורים להיות קרובים לנתונים כשאתם תריצו.

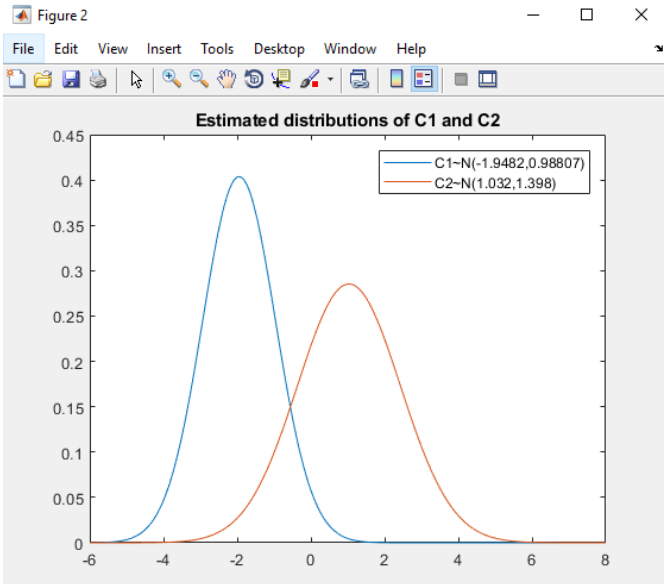
I. צור 300 סמפלים לכל קלאס



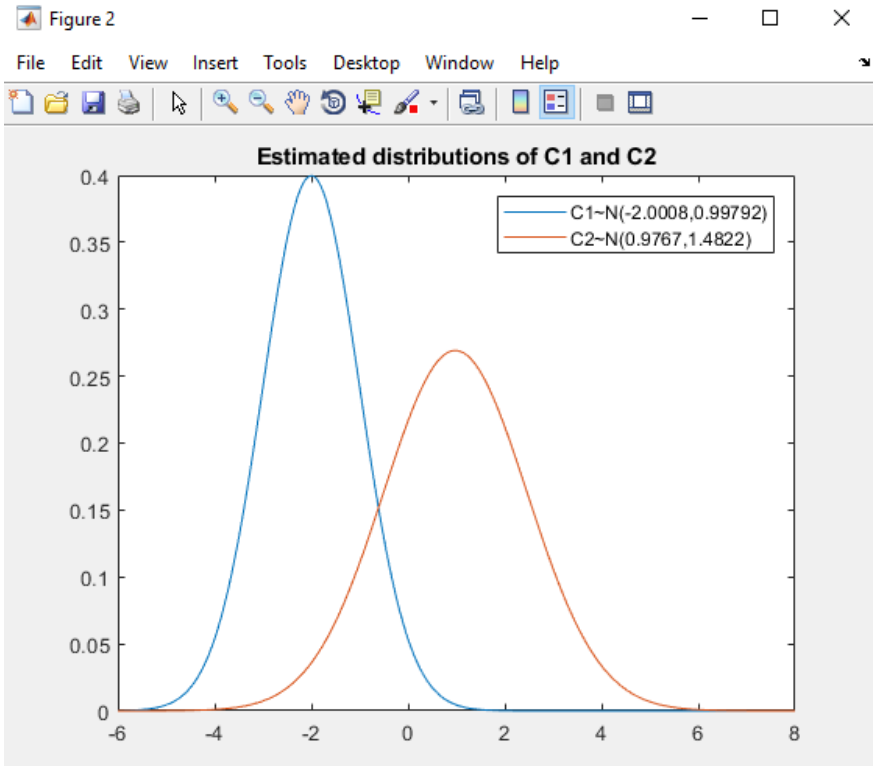

II. הערך את מיו וסטיית התקן והשווה לערכים האמיתיים

את הפרמטרים הערכנו לפי הגדרת מיו וSTD שראינו תחת MLE בהרצאה:

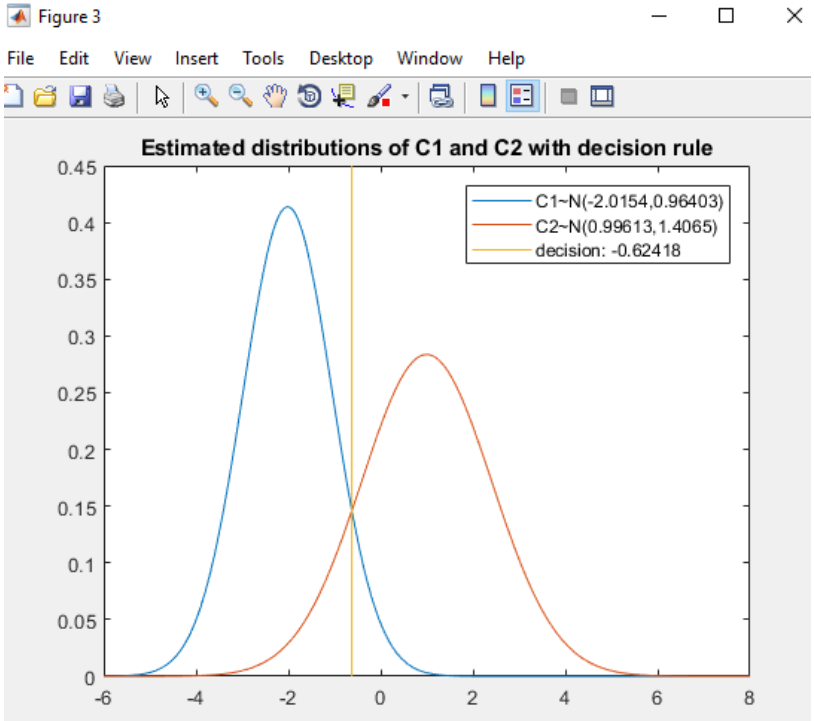
$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$



הערכת מיו וSTD עבור כל קלאס הניבה פעמוני גאוס די דומים להתפלגות האמיתית של הקלאס.
ההבדל נובע שהגרלנו את הDATA בצורה הסתברותית. ככל שיהיו יותר סמפלים, ההערכה של הפרמטרים תשאף להתפלגות האמיתית.
לדוגמא אם ניצור 3000 דגימות ונעריך את הפרמטרים, נקבל משהו שהרבה יותר קרוב להתפלגות האמיתית:




III. בהנתן ששגיאת הסיווג שווה לשני הקלאסים, מה חוק ההחלטה? ומה השגיאה עבור החוק הזה?



חוק ההחלטה הוא 0.62418- (עלות שגיאה זהה). פשוט השוונו בין הקלאס קונדישנס כמו בהרצאה:

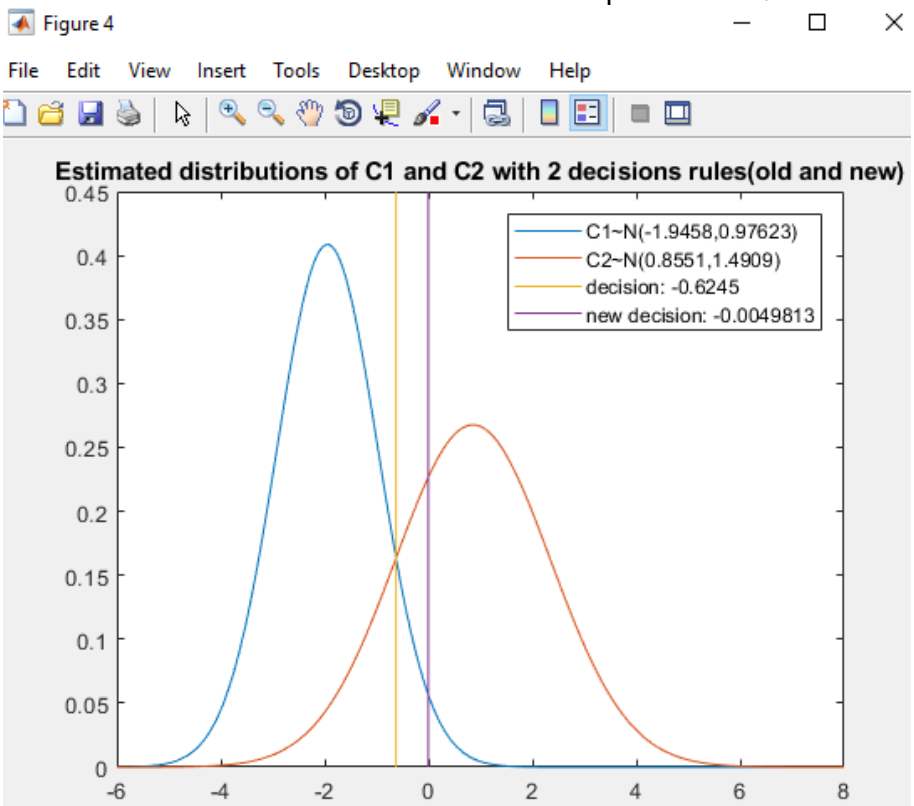
$$p(l \mid salmon) \ ? \ p(l \mid bass)$$

השגיאה היא בעצם כל הסמפלים של C1 שנמצאים מימין לקו ההפרדה + כל הסמפלים של C2 שנמצאים משמאל לקו ההפרדה

```
Error rate (same loss values)is 0.13833
```

IV. בהנתן שטעות בסיווג c1 עולה פי מטעות בסיווג c2, צייר על אותו גרף את חוק ההחלטה החדש.

אינטואיציה: אם c1 זה אדם חולה ו c2 הוא בריא, אז לראות c1 אדם חולה ולהגיד שהוא בריא עולה פי 4 מלהגיד לאדם בריא שהוא חולה. אם כך, עדיף למסווג להגיד לאדם שהוא חולה בכל פעם שיש לו ספק:




ולכן, הקו הסגול הוא הגבול החדש. כלומר המסווג מעדיף להגיד יותר שאנשים הם חולים.

$$\frac{P(l \mid b)\cancel{P(b)}\cancel{p(l)}}{\cancel{p(l)}P(l \mid s)\cancel{P(s)}} = \frac{P(l \mid b)}{P(l \mid s)} < \frac{\lambda_{bs}}{\lambda_{sb}}$$

C1 1 4 C2

כמו שניתן לצפות, השגיאה גדלה:

```
Error rate (cl loss is bigger)is 0.16
```

2. ממש את Naive Bayes Algorithm ודווח על אחוזי הסיווג.

השתמשנו בשני השקפים האחרונים בהרצאה של NB המתארים פסאודו קוד של האלגוריתם. קיבלנו אחוז הצלחה(בשבר) :

```
success rate on NBc: 0.96071
```

3. ממש את חלון פרזן עבור קלאסים B,A בעלי 16 מימדים. השתמש ב valid_data שלהם לבחירת גודל החלון ודווח על אחוזי הצלחת המסווג.

מימשנו את אלגוריתם חלון פרזן עם חלון גוסיאני כמתבקש בשאלה:

$$\frac{1}{n_j}\sum_{l=1}^{n_j}\exp\left(-\frac{\left\|x_l^j - x\right\|^2}{2\sigma^2}\right) \geq \frac{1}{n_i}\sum_{l=1}^{n_i}\exp\left(-\frac{\left\|x_l^i - x\right\|^2}{2\sigma^2}\right) \text{ for } \forall i \neq j$$

```
function scalar = gaus(sampleVecValue, testedVecValue, sig)
%       c = (1/sig*sqrt(2*pi));
    euclidinDistance = norm(sampleVecValue-testedVecValue);
    exponent = exp(-(euclidinDistance^2)/(2*sig^2));
    scalar = exponent;
end
```

הרצנו את valid_data עם סיגמות שונות וראינו שהתוצאות הטובות מתקבלות על סיגמא בין 0.2 ל2.3.

התוצאות הכי טובות על הטסט דאטה התקבלו בסיגמא 1.3,1.4,1.5,1.6 עם 99.718% סיווג:

```
succes rate(in fraction) for test data: 0.96901 0.1
succes rate(in fraction) for test data: 0.99155 0.2
succes rate(in fraction) for test data: 0.99155 0.3
succes rate(in fraction) for test data: 0.99155 0.4
succes rate(in fraction) for test data: 0.99155 0.5
succes rate(in fraction) for test data: 0.99155 0.6
succes rate(in fraction) for test data: 0.99155 0.7
succes rate(in fraction) for test data: 0.99155 0.8
succes rate(in fraction) for test data: 0.99437 0.9
succes rate(in fraction) for test data: 0.99437 1
succes rate(in fraction) for test data: 0.99437 1.1
succes rate(in fraction) for test data: 0.99437 1.2
succes rate(in fraction) for test data: 0.99718 1.3
succes rate(in fraction) for test data: 0.99718 1.4
succes rate(in fraction) for test data: 0.99718 1.5
succes rate(in fraction) for test data: 0.99718 1.6
succes rate(in fraction) for test data: 0.99437 1.7
succes rate(in fraction) for test data: 0.99437 1.8
succes rate(in fraction) for test data: 0.99155 1.9
succes rate(in fraction) for test data: 0.99155 2
succes rate(in fraction) for test data: 0.99155 2.1
succes rate(in fraction) for test data: 0.99155 2.2
succes rate(in fraction) for test data: 0.98873 2.3
best: 0.99718 1.3
done
```

4. ממש את KNN. השתמש ב Valid Data למצוא את K האופטמלי ודווח על תוצאות המסווג.

לאחר המימוש של KNN הרצנו את כל ה2000 K האפשריים. הנה תמונה של החמש הראשונים:

```
succes rate(in fraction) for valid data: 0.9825. K: 1
best K for valid data so far: 1 with success rate of 0.9825
succes rate(in fraction) for valid data: 0.9775. K: 2
succes rate(in fraction) for valid data: 0.9825. K: 3
succes rate(in fraction) for valid data: 0.9775. K: 4
succes rate(in fraction) for valid data: 0.9825. K: 5
succes rate(in fraction) for valid data: 0.9775. K: 6
```

קיבלנו תיקו משולש בין K=1,3,5.

אילו התוצאות עבור שלושתם:

```
succes rate(in fraction) for test data: 0.9975. with K: 1
succes rate(in fraction) for test data: 0.99. with K: 3
succes rate(in fraction) for test data: 0.9925. with K: 5
```

K האופטימלי הוא 1 עם הצלחה של 99.75%.